THE EXPRESSIVITY OF TURN-TAKING: UNDERSTANDING CHILDREN PRAGMATICS BY HYBRID CLASSIFIERS

Cristina Segalin[1], *Anna Pesarin*[1], *Alessandro Vinciarelli*[2], *Marco Cristani*[1]

[1] Università degli Studi di Verona, Strada Le Grazie 15, I-37134 Verona, Italy
[2] University of Glasgow, Computing Science Dept., Sir A. Williams Building, Glasgow G12 8QQ, UK

ABSTRACT

We analyze the effect of children age on pragmatic skills, i.e. on the way children manage the conversation dynamics. In particular, we focus exclusively on the turn-taking (who talks when and how much), reducing conversations as sequences of simple speech/silence periods. Employing a hybrid (generative + discriminative) classification framework, we demonstrate that such a simple signature is very informative, allowing to separate 22 "pre-School" conversations (between 3-4 years old children) and 24 "School" conversations (between 6-8 years old children), with 78% of accuracy. The framework exploits Steady Conversational Periods and Observed Influence Models as feature extractors, plus LASSO regression as feature selector and classifier. The generative nature of our method permits, as byproduct, to identify the pragmatic skills that better discriminate the two groups: notably, scholar children tend to have more frequent periods of sustained conversation, in a statistically significant way.

1. INTRODUCTION

The ability of sustaining a dialog depends on a tight timed coordination of speech, facial gestures, respiratory kinematics, bodily posture [1].

In this paper, we focus on the pragmatic skills that regulate the turn-taking (who talks when and how much), showing that they are related with the age of children; in particular, we designed a statistical framework that distinguishes prescholar and scholar conversations, starting from very simple patterns of silence and speech periods collected on dyads. As dataset, we consider a conversation set composed by 44 "pre-School" (3-4 years) and 48 "School" (6-8 years) italian subjects.

The proposed approach is based on an hybrid classification framework [2], where training data is initially learned by generative models; after that, the parameters of the models are employed as features by a discriminative approach.

In our case (see Fig. 1), we firstly extract Steady Conversational Periods (SCP) [3] from conversation recordings; they are low-level cues, which essentially assume a dyad as two coordinated Markov chains (one for each participant): whenever a turn starts, finishes or it is interrupted, a couple of SCPs

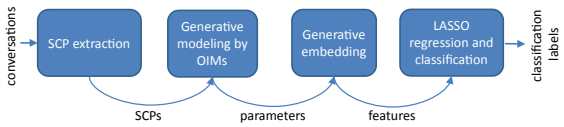

Fig. 1. Scheme of the proposed approach.

(one for each subject) are instantiated. This enforces synchronization between the Markov chains, allowing to treat them as a single stochastic process, here captured by an Observed Influence model (OIM) [4].

For each dyad, we learn an OIM: once all the training dyads have been processed, the related OIMs can be transformed into features by Generative Embedding [2]; the idea is that the parameters of the OIM generative models can be seen as features, projected in a metric space. Here, discriminative approaches are trained to reach high classification scores.

In this work, we embed a feature selection phase in the classification step, adopting LASSO regression as feature selector and discriminative classifier. With LASSO, a restricted pool of features is automatically selected and employed to separate the two classes. This amounted to a 78% of Leave-One-Out classification accuracy.

In addition, we perform statistical analysis of the features selected by LASSO, discovering significant differences among the two classes, that highlight the tendency of the scholar population to have a more sustained dialogs, with shorter and more frequent turns occurrences.

In the rest of the paper, Sec. 2 provides a brief overview of related work, Sec. 3 illustrates the proposed methodology, Sec. 4 reports on experiments and results, and Sec. 5 draws some conclusions.

2. RELATED WORK

The computing literature proposes a large number of works where pragmatics related measurements (e.g., total speaking time, statistics of turn length, prosody, voice quality, etc.) are shown to be the evidence of social and psychological phenomena (see [5] for an extensive survey). Examples include the work in [6], where a dialogue classification system discriminates three kinds of meetings using probability transi-

tions between periods of speech and silence, the experiments in [7], where features based on talkspurts and silence periods (e.g., the total number of speaking turns and the total speaking length) model dominance, the approach of [8], where intonation is used to detect development problems in the early childhood, and the work in [9], where prosody analysis allows the identification of language impaired children.

3. THE APPROACH

This section follows the scheme of Fig. 1, giving a short explanation of the first three modules, focusing more on the LASSO regression and classification.

3.1. Steady Conversation Period Extraction

The first step of the approach operates directly on the raw conversations, extracting the *Steady Conversation Periods* (SCP) [3]: at every instant, every conversation participant i is in a state $k_i \in [0, 1]$, where 0 corresponds to being silent and 1 to speak, and $i = 1, \ldots, C$, where C is the total number of conversation participants[1].

A SCP is the time interval between two consecutive state changes (not necessarily of the same participants). Hence, there is a sequence of SCPs for each participant i: $\{(d(n), k_i(n)\}$, where $d(n)$ is the duration of the SCP and $k_i(n)$ is the state of speaker i in SCP n. Length of the sequence and duration $d(n)$ of every sequence element are the same for all participants because the SCP changes whenever any of the participants changes state.

Overall, the extraction of the SCPs corresponds to a segmentation of the conversation into intervals during which the configuration (who talks and who is silent) is stable. In order to take into account different durations while keeping a low number of states in the Observed Influence Model (see below), the durations $d(n)$ are grouped into $D = 2$ classes (*short* and *long*) by an unsupervised Gaussian clustering performed over a training dataset. This creates $D \times 2 = 4$ different types of SCP: *long silence*, *long speech*, *short silence*, *short speech*.

Supposing V conversations, this step provides V sequences of SCPs, where each sequence reports the SCPs of both the dialog participants.

3.2. Generative Modeling by OIMs

The Observed Influence Model (OIM) [4] is a generative model for C interacting Markov chains. For a chain i ($i = 1, \ldots, C$), the transition probability between two consecutive states $S_i(t-1)$ and $S_i(t)$ is:

$$P(S_i(t)|S_1(t-1), \ldots, S_C(t-1)) = \\ = \sum_{j=1}^{C} {}^{(i,j)}\theta P(S_i(t)|S_j(t-1)) \quad (1)$$

where $1 \leq i, j \leq C$, ${}^{(i,j)}\theta \geq 0$, $\sum_{j=1}^{C} {}^{(i,j)}\theta = 1$, and $P(S_i(t)|S_j(t-1))$ is the probability of chain i moving to state $S_i(t)$ at step t when chain j is in state $S_j(t-1)$ at step $t-1$. An OIM can be defined as $\lambda =< A^{(i,j)}, \pi, \theta >$ $(1 \leq i, j \leq C)$ where $A^{(i,j)}$ is the matrix such that $A_{kl}^{(i,j)} = P(S_i(t) = l|S_j(t-1) = k)$, π is a $C \times L$ (L is the total number of states) matrix such that $\pi_{ik} = P(S_i(1) = k)$ and θ is a $C \times C$ weights matrix where $\theta_{ij} = {}^{(i,j)}\theta$. In our case, we have dialogic conversations, i.e., $C = 2$; we have also $L = 4$ states corresponding to the four kinds of SCPs. Therefore, having V conversations, we learn V OIMs, $\{\lambda_v\}$, $v = 1, \ldots, V$.

3.3. Generative Embedding

Roughly speaking, the generative embedding (GE) is a sort of feature extraction that consists in the use of generative model parameters as features, so that a further step of (discriminative) classification can be performed [2].

In our case, we extract the transition matrices $A^{(i,j)}$.[2] More into detail, we collapse transition probabilities as follows:

$$\tilde{A}_{kl}^{(i,j)} = \frac{1}{2}\left(A_{kl}^{(i,j)} + A_{kl}^{(j,i)}\right) \text{ if } i \neq j \quad (2)$$

$$\tilde{A}_{kl}^{(i,i)} = \frac{1}{2}\left(A_{kl}^{(1,1)} + A_{kl}^{(2,2)}\right) \quad (3)$$

It basically extracts inter and intra probability values, averaging over the different speakers, reaching thus invariance with respect to the speakers order. At the end, avoiding repeated values, the feature vector ψ_v for each model λ_v has size $C \times L^2 = 32$.

3.4. Lasso Regression

Given the pool of V features vectors, we perform a sparse regression analysis using Lasso [10], for feature selection and classification purposes. Lasso is a general form of regularization in a binary regression problem. Let suppose that the V features vectors represent the training data. In the simple linear regression problem every training sample ψ_v is associated with a target variable y_v, that in our case is the class label $\{1; -1\}$ Then, we can express the target variable as a linear combination of the generative features:

$$y_v = \mathbf{w}^T \psi_v \quad (4)$$

The standard least square estimate calculates the weight vector $\mathbf{w}$ by minimizing the error function

$$E(\mathbf{w}) = \sum_{v=1}^{V} \left(y_v - \mathbf{w}^T \psi_v\right)^2 \quad (5)$$

The regularizer in the Lasso estimate is simply expressed as a threshold t on the L1-norm of the weight $\mathbf{w}$, *i.e.*, $\sum_j |w_j| \leq t$;

[1] Silence/speech separation has been achieved by manual annotation.

[2] We found that considering the coefficients ${}^{(i,j)}\theta$ or the initial state probabilities does not help in the classification.

the term t acts as a constraint that has to be taken into account when minimizing the error function.

By doing so, it has been proved that (depending on the parameter t)[3], many of the coefficients w_j become exactly zero [10]. Since each component w_j of the weight vector weighs a different feature of the feature vector ψ_v (*i.e.*, a transition probability), it is possible to understand which transitions are the most discriminative for the classes at hand. In particular, by looking at the high absolute values in $\mathbf{w}^{(n)}$, we can observe the most important features for the classification: the higher the value, the more important the feature.

Given a test sequence ψ_{test}, obtained by learning an OIM model λ_{test} on a test conversation, and multiplying it by $\mathbf{w}$, it outputs a score β_{test}. Its sign indicates the winning class.

4. EXPERIMENTS

We use our hybrid classification framework to analyze the effects of age on pragmatic skills for children between 3 and 8 years old. The analysis is organized in three parts: 1) a quantitative analysis of the dataset, 2) the review of the LASSO classification results and 3) a psychological interpretation of the features selected by the LASSO classifier.

4.1. The Data

The corpus used for the experiments includes 46 dyadic conversations between Italian children (92 subjects in total). The corpus is split into two parts: 22 conversations involve 3-4 years old children, named *pre-School* (pS). The other 24 conversations include 6-8 years old children, named *School* (S). All the conversations hold between different subjects, considered once. The experimental setting corresponds to a *controlled observation*: the children sit close to one another and fill an album, in a situation not particularly different from their everyday experience. The average duration of the conversations is 15 minutes and 31 seconds for pS children and 15 minutes and 21 seconds for S children. The conversations have been recorded with an unobtrusive Samsung Digital Camera *34*×.

Data was manually processed independently by two different annotators, in order to perform error-free source separation; as silence periods we considered segments that don't contain sounds or sounds like cough, sneezing, ambient noise. As speech, we considered all other segments that contain verbal sounds. Silences shorter than 600 ms have been considered part of a *speech* segment.

4.2. Quantitative Analysis of the Dataset

After the extractions of the SCP, we analyze the average percentage of silence and speech SCPs for pS and S conversations, see Table 1. The table shows no significant differences between the two classes of conversations; in addition, a very low standard deviation for both the classes (0.008 for the pS and 0.01 for the S) indicates a strong similarity among the conversations.

[3]In this work, the t parameter has been chosen by cross-validation.

Class	Silence SCP	Speech SCP
pS	74%	26%
S	72%	28%

Table 1. Amount of silence and speech SCPs for each class.

Class	Short Sil.	Long Sil.	Short Sp.	Long Sp.
pS	56%	15%	22%	7%
S	58%	12%	24%	6%

Table 2. Average percentages for short and long SCPs.

		PreS		**S**	
Method	Acc.	Prec.	Rec.	Prec.	Rec.
Histogram-based	61%	57%	72%	67%	50%
Our approach	**78%**	**71%**	**91%**	**89%**	**67%**

Table 3. Classification results.

The clustering of the SCP durations into four states (S_1 = *short silence*, S_2 = *long silence*, S_3 = *short speech* and S_4 = *long speech*) produces the following duration statistics - means (dev. std) -: $S_1 = 1.37s$ (1.07), $S_2 = 19.36s$ (29.7), $S_3 = 1.3s$ (0.74), $S_4 = 4.10s$ (2.69). Given this quantization, the proportions of the four states in the two classes are shown in Table 2. In this case, we can note that short silence and short speech SCP are slightly more frequent in the scholar class. At this point, one can suppose that the duration information only should discriminate the two classes. We will come back on this point in the next section.

4.3. Classification and Parameters Analysis

The classification protocol is based on Leave-One-Out cross validation. At each run of the cross-validation, the training set composed by $V-1$ elements is processed by LASSO, producing a weight array $\mathbf{w}_v$ which serves to classify the $V-$th test element.

After the cross-validation, the resulting accuracy, precision and recall for each class is reported in Table 3.

As comparative test, we consider solely the silence and speech SCP durations, without accounting for turn-taking information. For each training conversation, we calculate the histograms of the SCP silence and SCP speech durations, both of 16 bins: it is worth noting that here we do not consider the *quantized* SCP durations, but their original values prior to the clustering; this way, the original speech and silence durations were taken into account. In addition, the binning of the histograms is exponential, with denser bin intervals at short durations: this allows to better account the large amount of short SCP durations. After that, we concatenate the histograms

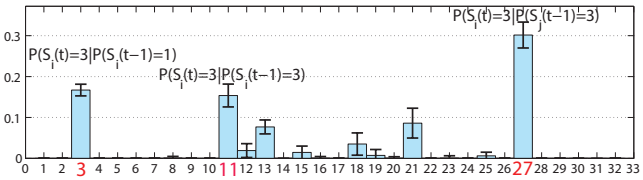

Fig. 2. Feature selection by LASSO. Best viewed in color.

obtaining V $32-$dimensional vectors, feeding them into the Lasso classifier, employed with the same classification protocol for the generative embedded data. The performances of this method, dubbed here "Histogram-based" are reported in Table 3. As visible, the contribute given by the turn-taking information is strongly informative, and in the next section we will see in which respect.

4.4. Psychological Interpretation of the Features

At the end of the cross-validation cycle, we have V weight vectors $\{\mathbf{w}_v\}$, one for each training/testing partition. Averaging over the absolute values of their values we get the mean weights associated to each feature; the variance is also calculated, and shown in Fig. 2.

As visible, many features have been set to 0 by LASSO, meaning that they are not useful to discriminate the two conversation classes. To get more insight, we analyze the values for all the features, looking for inter-class statistical difference. In particular, we apply the Two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test, which fits well the data cardinality at hand. The significantly different features are the 3,11,27 (they have been all selected by LASSO), with p-value 5% (depicted in red on Fig. 2). Feature f_3 indicates the probability that a subject utters a short sentence after he was silent for a short time; features f_{11} and f_{27} indicate the probability of having a short speech segment after another short speech period of the same subject, or uttered by the other interlocutor, respectively. This indicates the presence of overlapping speech or (less frequently) an alternation of speech periods without pauses inside. All these probabilities are higher in the case of the scholar class of an average of 0.03, indicating that S subjects seem to keep a higher conversational rhythm compared to pS subjects.

5. CONCLUSIONS

This paper offers a novel study of how effectively turn taking markers can discriminate the age of children. The use of Steady Conversational Periods, fed into hybrid classifiers, allowed to finely separate classes of pre-scholar and scholar conversations, explaining actually how the two classes are different: scholar children tend to have more frequent periods of sustained conversation. This study promotes many future developments, for example the investigation of intra class differences in the set of scholar or prescholar subjects; more importantly, this approach may lead to the definition of a clinical semeiotics able to individuate automatically pragmatic language impairments.

6. REFERENCES

[1] S.K. Scott, C. McGettigan, and F. Eisner, "A little more conversation, a little less action. candidate roles for the motor cortex in speech perception," *Nat. Rev. Neurosci.*, vol. 10, pp. 295–302, 2009.

[2] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "Free energy score spaces: Using generative information in discriminative classifiers.," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1249–1262, 2012.

[3] M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina, and V. Murino, "Generative modeling and classification of dialogs by a low-level turn-taking feature," *Pattern Recogn.*, vol. 44, no. 8, 2011.

[4] B. Clarkson S. Basu, T. Choudhury and A. Pentland, "Towards measuring human interactions in conversational settings," in *IEEE Int'l Workshop on Cues in Communication (CUES 2001)*, Hawaii, CA, 2001.

[5] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signals, their function, and automatic analysis: a survey," in *ICMI '08*, 2008.

[6] K. Laskowski, "Modeling vocal interaction for text-independent classification of conversation type," in *Proc. SIGdial*, 2007, pp. 194–201.

[7] H. Hung, Y. Huang, G. Friedl, and D. Gatica-Perez, "Estimating the dominant person in multi-party conversations using speaker diarization strategies," in *ICASSP*, 2008.

[8] A. Mahdhaoui, M. Chetouani, R.S. Cassel, C. Saint-Georges, E. Parlato, M.C. Laznik, F. Apicella, F. Muratori, S. Maestro, and D. Cohen, "Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents," *International Journal of Methods in Psychiatric Research*, vol. 20, pp. e6–e18, 2011.

[9] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for the prosodic assessment of language impaired children," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 19, no. 5, pp. 1328–1342, 2011.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.